DOCUMENT RESUME

ED 164 610 TM 008 178

AUTHOR Merz, William R.; Rudner, Lawrence M.
TITLE Bias in Testing: A Presentation of Selected Methods.
PUB DATE Mar 78
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS Analysis of Covariance; Analysis of Variance; Complexity Level; *Evaluation Criteria; *Evaluation Methods; Factor Analysis; Item Analysis; Predictive Validity; *Predictor Variables; Scores; *Test Bias; *Test Items; Test Selection
IDENTIFIERS Chi Square; Item Characteristic Curve Theory; Item Discrimination (Tests)

ABSTRACT

A variety of terms related to test bias or test fairness have been used in a variety of ways, but in this document the "fair use of tests" is defined as equitable selection procedures by means of intact tests, and "test item bias" refers to the study of separate items with respect to the tests of which they are a part. Seven different operational definitions of the fair use of tests are described; distinctions made between those applied to fairness for individuals who are members of special groups, and those applied to fairness for groups but not their individual members. All seven approaches use the regression model. One method also requires the use of expected utilities. Various methods are described for investigating test item bias. Both classical test theory item analysis and latent trait item characteristic curve approaches are mentioned. Tests for bias included analysis of variance, chi-square, factor analysis, and arbitrary confidence bands. Distractor responses and item-test point-biserial correlations have also been considered. All of these methods are described briefly, but no attempt was made to evaluate them. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED164610

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Bias in Testing: A Presentation of Selected Methods

William R. Merz, Ph.D.
Associate Professor, Department of Behavioral Sciences in Education
California State University, Sacramento

and

Lawrence M. Rudner, Ph.D.
Research Specialist, Model Secondary School for the Deaf
Gallaudet College

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
William R. Merz
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A paper presented as part of a Co-Sponsored Symposium, Empirical Evaluations of Various Approaches for Identifying Biased Test Item(s), at the Annual Meeting of the American Educational Research Association and the National Council for Measurement in Education, Toronto, Canada, March 1978.

IM008 178

Printed in U.S.A.

The issue of collecting fair information on the performance of members of identifiable groups is a major problem in the construction and use of tests in schools, government, and industry in the United States of America. Interest in this issue has broadened beyond the boundaries of the United States to other countries with multiethnic and multilingual populations. These issues have been addressed under various labels; however, the two most frequently used are test fairness and test bias. The two terms often have been used synonymously. It is the contention of the present reviewers (Merz, 1976; Rudner, 1977c) that synonymous use has led to major confusion about methods and loss of focus on the questions being addressed.

One set of methodologies is used for situations in which an intact test is administered to members of different groups and these groups obtain different mean total scores. Under this condition the usual goal of measurement is to provide data for selection of applicants. The issue is using the test to predict later success in a fair and equitable manner. Here, one of the approaches to regression analysis is applied with the test of interest as a predictor and some external measure of success as the criterion.

Another set of methodologies involves the identification of items which systematically differentiate among members of a group. To date these methodologies have not used an external criterion of success and have focused on single items from a pool which constitutes

or will constitute the intact test. The focus of these efforts is usually the construction of a measure which assesses a content domain without introducing systematic variance attributable to factors other than those which are the intended object of measurement.

The purpose of this paper is to describe the methods used to investigate the presence of what has been labeled as "test bias." In order to make a clear distinction between the two methodologies, the examination of intact tests for equitable selection will be treated under the topic Fair Use of Tests; while examination of items within a test or an item pool for systematic performance differences among groups will be described under the heading Test Item Bias.

Fair Use of Tests

This type of investigation is of interest to test users who need to know the accuracy of test information. Seven approaches to determining fairness in selection are reviewed here. The seven are all regression approaches; that is, they attempt to predict from a selection or placement instrument to a criterion of success. Each uses a correlation-prediction model, but each differs in the way the criterion cut-off score is adjusted to yield fair estimates of success. Thus, each method assumes that there is a valid, reliable, and unbiased criterion measure for members of a given group. Further, the other assumptions of regression models also pertain--bivariate normality and homogeneity. The first assumption is absolutely necessary; if the criterion is not valid, reliable, and unbiased, then, the prediction method fails. The assumptions of normality and

homogeneity systematically affect the magnitude of the correlation coefficient and, thereby, influence the accuracy of prediction. Of course, a sufficient number of examinees must be available to compute stable correlations, or the method is unreliable.

The first method, labeled the regression model by Petersen and Novick (1976), was described by Cleary (1968). It defines a test as fair if there are no consistent non-zero errors of prediction for members of each subgroup of the population. This relationship is described by this equation:

$$y^* = \alpha_1 + \beta_1 x_1^* = \ldots = \alpha_g \beta_g x_g^*$$

where α_i represents the intercept

β_i represents the slope, and

x_i^* represents the predictor pass score for subpopulation π_i $(i = 1,\ldots,g)$.

Here a different regression equation is calculated for each subgroup. Corrections are made because of differences in mean values of X and Y among subgroups. However, only one acceptable criterion score is used. Hence, Darlington (1971) views this situation as $r_{cx} = r_{cy}/r_{xy}$; that is, the correlation between group membership and the predictor is equal to the ratio of the correlation between group membership and the criterion to the correlation between predictor and criterion. The focus of this method is on fairness to the individual rather than on fairness to the group. It is the most widely used approach to fair selection.

The second method was described by Thorndike (1971) and developed by Cole (1973); it was called the constant ratio model by Petersen and Novick (1976). A test is fair if it identifies applicants for selection in such a way that the ratio of the proportion selected to the proportion successful is the same in all subpopulations. Here the relationship may be described as:

$$R = \frac{\text{Prob}(X \geq x_1^* | \pi_1)}{\text{Prob}(Y \geq y^* | \pi_1)} = \ldots = \frac{\text{Prob}(X \geq x_g^* | \pi_g)}{\text{Prob}(Y \geq y^* | \pi_g)}$$

where R is a fixed constant for subpopulations and x_i^* represents the predictor cut-off score for subpopulation $\pi_i (i = 1, \ldots, g)$.

This method focuses on fairness to the group rather than on fairness to the individual. It requires, in addition to the general assumptions listed earlier that a constant ratio of success is reasonable in all subgroups; here, $r_{cx} = r_{cy}$ (Darlington, 1971). This approach is used where equity between groups is the central consideration and in situations where the differences between the means for predictors is different from the differences between means for the criterion. Emphasis is on false successful and false unsuccessful predictions, as well as on accurate predictions Petersen and Novick (1976).

A third approach was proposed by Einhorn and Bass (1971); it was labeled the equal risk model by Petersen and Novick (1976). It defines a test as fair when all persons selected are predicted to be above a specific minimum point on the criterion with a specified degree of confidence. In this case,

$$Z = \text{Prob}(Y \geqslant y^* \mid X = x_1^*, \pi_1) = \ldots$$

$$= \text{Prob}(Y \geqslant y^* \mid X = x_g^*, \pi_g)$$

where Z is a fixed constant probability for all subpopulations π_i $(i = 1, \ldots, g)$

x_i^* represents the predictor cut-off score for a subpopulation, and

Y^* represents the criterion cut-off score.

This is accomplished by adjusting the criterion passing score with a confidence band, so that

$$Y' = Y_c - Z_p \, (s_{y \cdot x})$$

where Z_p is a z-score which can be designated by the desired degree of risk,

Y_c is the criterion cut-off, and

$s_{x \cdot y}$ is the standard error of estimate.

This approach allows separate cut-off points for each subpopulation if the standard errors for subpopulations are different. Where standard errors for each subpopulation are equal, the results reduce to the situation described for the regression model. It, too, focuses on the group rather than the individual. Other assumptions are similar to the regression model. In addition, it must be logical to expect the probability of success in each group to be equal.

Darlington (1971) suggested a model which would replace the concept of cultural fairness with another which he labels cultural optimality; hence, it was called the culture modified criterion

approach by Petersen and Novick (1976). Darlington defines a test as culturally optimal when; (1) a subjective policy level question is answered concerning the optimum balance between performance and cultural factors, and (2) an empirical relationship between a test and a culture modified variable (Y-kC) is established.

Here, $(Y - kC) = \alpha_i + \beta_i x_i^*$

where Y is the criterion

k is a constant subjective value judgment on the part of the decision maker,

C denotes an applicant's group membership

α_i is the intercept,

β_i is the slope, and

x_i^* is the predictor pass score for subpopulations π_i $(i = 1,\ldots,g)$.

The criterion score is adjusted by a predetermined amount based upon group membership. Here, in addition to the other assumptions inherent in a regression approach, one must see some value in selecting members of a subpopulation and that value must be translated into the constant which adjusts the criterion. The process of adjusting criterion scores is open and may be debated publicly.

Cole (1973) proposed a fifth method labeled the conditional probability model by Petersen and Novick (1976). In this model, a test is regarded fair if, given satisfactory criterion performance, individuals have the same probability of selection regardless of group membership.

Here, $P = \text{Prob}(X \geq x_1^* \mid Y \geq y^*, \pi_1) = \ldots$

$= \text{Prob}(X \geq x_g^* \mid Y \geq g^*, \pi_g)$

where P is a fixed constant for subpopulations $\pi_i (i = 1, \ldots, g)$ and

x_i^* represents the predictor pass score for subpopulation π_i. This model looks for equity to the group. The emphasis is on false unsuccessful predictions as well as on accurate predictions (Petersen and Novick, 1976). In addition to the other assumptions mentioned earlier, it must be reasonable to expect that all groups perform equally well on the criterion.

The sixth model was proposed by Linn (1973) and defines a test as fair if all applicants who are selected are guaranteed an equal, or fair, chance of being successful regardless of group membership. This model was labeled the equal probability model by Petersen and Novick (1976).

Here, $Q = \text{Prob}(Y \geq y^* \mid X \geq x_1^*, \pi_1) = \ldots$

$\text{Prob}(Y \geq y^* \mid X \geq x_g^*, \pi_g)$

where, Q is a fixed constant for all subpopulations $\pi_i (i = 1, \ldots, g)$

x_1^* represents the predictor cut-off score for subpopulation π_i

It, too, seeks equity for the group. It emphasizes false unsuccessful predictions as well as accurate predictions (Petersen and Novick, 1976). In addition to the other assumptions mentioned earlier, it must be reasonable to expect all groups to perform equally well on the predictor.

The last model to be reviewed was proposed by Gross and Su (1975) and was labeled the threshold utility model by Petersen and Novick (1976). It states that a test is fair if an individual from a subpopulation is selected when his/her predicted score reaches a specific minimum point on the criterion which has been modified in such a way that the expected utility of the selection process is a maximum. Here the utility of the selection process is found with

$$\varepsilon[u(0)] = \sum_{i=1}^{2} P_i \sum_{j=1}^{4} u(0_j|\pi_i) \text{ Prob } (0_j|\pi_i)$$

where P_i is the proportion of the combined applicant population (π_1 and π_2) who are members of the subpopulation.

This assumes that four outcomes are possible:

0_1: $X \geq x_i^*$ $Y \geq y^*$ An applicant is accepted and is successful

0_2: $X \geq x_i^*$ $Y \geq y^*$ An applicant is rejected but would have been successful

0_3: $X \geq x_i^*$ $Y \geq y^*$ An applicant is rejected and would have been unsuccessful

0_4: $X \geq x_i^*$ $Y \geq y^*$ An applicant is accepted and is unsuccessful

Utilities usually differ for each subpopulation, and regression equations may differ, too. The method escapes the difficulty of emphasizing false successful and false unsuccessful predictions by seeking a public statement of utilities and, then, maximizing the likelihood of that utility for a given group.

Test Item Bias

This type of investigation is of interest to test developers because it assists them in devising valid, cross-culture fair items and provides a framework for constructing better tests in subsequent efforts. Six approaches are reviewed here.

Analysis of Variance Approaches

Cardall and Coffman (1964) suggested a method of identifying bias using an analysis of variance framework which incorporates test items and group membership as main effects. Bias is defined as a significant item by group interaction, that is, the presence of items which are relatively more difficult for members of one culture group than another. In order to meet the homogeneity of variance assumption of the analysis of variance, Cardall and Coffman transformed the within group item difficulties with an arcsin transform.

Plake and Hoover (1977) extended the technique to allow for the identification of individual items. The interaction contrast for each item within each group takes the form:

$$\psi_{ij} = \theta_{ij} - \theta_{i\cdot} - \theta_{\cdot j} + \theta_{\cdot\cdot}$$

where θ_{ij} is the arcsin transformed item difficulty for the ith and the jth group.

The error variance is

$$\sigma_{\psi_{ij}} = \frac{q-1}{2q\,\tilde{n}_j}$$

where q is the number of items

$\tilde{n}_j$ is the harmonic mean of the number of subjects in the jth group.

A simultaneous significant test, such as Bonfersoni's procedure, can then be used to identify individual items which appear to be biased.

Cleary and Hilton (1968) employed a three factor, mixed model analysis of variance to determine whether or not items within a test were biased; again defining bias as items by group interaction. In their analysis, race and socioeconomic status were considered fixed variables; while persons and items were considered random. Socioeconomic status levels were nested within race to avoid assuming that the levels were comparable across the races.

Other examples of this approach can be found in Eagle and Harris (1969), Hoepfner and Strickland (1972) and Jenson (1973). It should be noted that these authors and Cardall and Coffman did not incorporate an arcsin transform.

Transform Item Difficulties

The transformed item difficulties approach, providing for a visual examination of item by group interaction effects, was probably first described by Thurston (1925) in connection with his method of absolute scaling. Of the approaches, this method appears to be one of the best known. It has been advocated and used frequently by Angoff (1972), Angoff and Ford (1974), and Angoff and Modu (1973), and others (Green and Draper, 1972; Jensen, 1973; Hicks, et al. 1976; Strassberg-Rosenberg and Donlon, 1975; Echternacht, 1974; and Rudner, 1978). Further, the approach has appeared in at least one measurement textbook (Anastasi, 1976, pp. 222-226).

In this method, indices of item difficulty; i.e., p-values--are obtained for two different groups on

> a number of items. Each p-value is converted to a normal deviate and the pairs of normal deviates, one pair for each item, are plotted on a bivariate graph, each pair represented by a point on the graph. (Angoff, 1972, p. 1)

The plot will generally be in the form of an ellipse. A 45° line, passing through the origin, provides a theoretical regression indicating the absence of bias. Items greatly deviating from this line may be regarded as exhibiting an item by group interaction. Relative to the other items, deviant items are especially more difficult for members of one group than the other. Assuming both groups received similar instructions, such items would appear to represent different psychological meanings for the two groups of examinees.

Since the intent is to make comparisons of between-group differences in item difficulty, it is necessary to transform the proportion passing an item to an index of item difficulty which constitutes at least an interval scale. This is accomplished by expressing each item p-value in terms of within-group deviations of a normal curve (see Guilford, 1954, pp. 418-419). Any linear transformation of the item z-score will meet such a requirement. One such transformation has been Delta values (4z + 13).

The distance of an item point to the line,

$$d = (z_1 - z_2)/\sqrt{2}$$

where z_j is the transformed item difficulty for group j, and serves to indicate the degree of item bias. Items which are "greatly deviating" from the line are identified by a traditional or nontraditional method of outlier or residual

analysis. One method is to place confidence limits on the line by using a multiple of the standard error of estimation. An alternate approach, adopted by Strassberg-Rossenberg and Donlon (1975) and Hicks, et al. (1976) involves computing the standard deviation of the residuals and classifying as biased those items deviating by greater than 1.5 standard deviation units. Rudner (1978) has employed a fixed item-regression line distance of .75 z-score units.

Echternacht (1974) also began with item difficulties which were transformed to delta values. Differences in transformed item difficulties were computed for each pair of groups, and these differences were plotted on normal probability paper. A-ditionally, a line was plotted to represent a hypothetical normal distribution with the obtained mean and standard deviation of the difference between pairs as parameters. Confidence bands constructed around this line represent the area outside of which biased items would fall.

Correlation Approaches

These approaches examine the point biserial correlation coefficients between item performance and total score. Ozenne, Van Gelder, and Cohen (1974) coupled a graphing method with the point biserial correlation approach. First, item difficulty levels were plotted using one group as a reference against which the other groups were plotted. Items were arranged in order of difficulty for the reference group from most difficult to least difficult; item numbers were plotted along the ordinate and item difficulty, along the abscissa. In this case a

publisher's national standardization sample was used as the reference group against which a minority sample was plotted. Visual examination of the plots revealed item by group interactions when the uniformity of the shapes of the curves was disturbed. The magnitudes of differences were not the concern; rather the deviation from the shape of the reference curve was noted. Then, point biserial correlations between item scores and total score were computed for each group that was to be compared. Correlations were compared to identify items which for a particular group did not contribute to total score; that is, items with a low item-total score correlation for a specific group were examined for bias. Items were identified as potentially biased by expert judgment based on the results of the two methods of analysis.

Green's strategy was used in standardizing the Comprehensive Tests of Basic Skills, Form S (Green, 1976; CTB/McGraw-Hill, 1974). Again, point biserial correlations were computed for each group on each item; any item having a correlation of less than .20 for any group was deleted. Green offered as evidence for the effectiveness of this strategy that fewer point biserial correlations fell below .20 for blacks in the standardization data.

Factor Analytic Approaches

In factor analysis, underlying factors (i.e., dimensions or traits) are hypothesized and the correlations of each variable with the hypothesized factors are computed. In an achievement test, each item is treated as a variable. Such an

analysis could be conducted twice using examinees from two different cultural backgrounds. Ideally, the two separate groups of examinees would yield similar sets of item-trait correlations (factor loadings). Different sets of factor loadings would indicate that the two groups are not responding to the items in the same manner. Such a test would be considered biased in that it would appear to measure different traits across groups. The items exhibiting the most bias would then be those with the largest differences in factor loading.

The general model for this type of factor analysis is

$$\underline{y} + A\underline{f} + \underline{e}$$

where, $\underline{y}$ is a vector of subject responses

A is a matrix of factor loadings

$\underline{f}$ is a vector of factor variables (locations)

$\underline{e}$ is a vector of residual or error terms

From $\underline{y}$, values of A, $\underline{f}$, and $\underline{e}$ are determined.

Green and Draper (1972) and Green (1976) suggest an intergroup factor analysis model based on the inter-battery factor analysis approach offered by Tucker (1958). In this intergroup model, the item variance is partitioned into: (1) factors common to each subgroup; (2) factors specific to subgroups, and (3) residual or error variance. With this model one can determine the proportion of item variance accounted for by a given subgroup. An item, then, is unbiased when this proportion is small and biased if a large proportion of variance is attributable to culture-specific sources.

Merz (1973, 1976) developed an alternate approach which

incorporates factor scores and anlysis of variance. The item intercorrelation matrix is computed for subjects pooled across groups. The matrix is reduced with Principal Components Analysis, employing theScree technique to determine the number of factors to be extracted. The factor matrix is then rotated orthogonally to simple structure, and factor scores derived from the rotated matrix.

An analysis of variance is then conducted on each set of factor scores using multiple group memberships as independent variables and factor scores for each vector as the dependent variable. Item bias is defined as a major loading on a factor with a significant F ratio on a main effect or on an interaction.

Distractor Response Analysis

Veale and Foreman (1975, 1976) recommend investigating the distractor response distribution for various cultural groups in an approach not dependent on total test scores. Should one group be overly attracted to a particular distractor in comparison to a second group, there may be a biasing characteristic of the item attracting them away from the correct response. Bias is defined as characteristics of an item which causes a distortion in the item p-value for a cultural group.

Consider the choice distribution illustrated in Table 1. Observed frequencies appear in the cells and expected frequencies appear in the upper right hand corner of each cell. A disproportionate number of members of Group 2 were attracted to Distractor 1 (the response frequencies can be shown to be disproportionate by the use of a chi-square test). It is

argued that some characteristic of Distractor 1 caused a substantial number of members of Group 2 to select this distractor over the correct alternative. Hence, some characteristics of the item may have caused a distortion in the group p-value.

Table 1

A Hypothetical Item Distractor Choice Distribution

Frequency of Selection

	Distractor 1	Distractor 2	
Group 1	60 40	40 60	100
Group 2	60 80	40 20	100
	120	80	200

Maw (1977) has developed an approach based on the work of Ku and Kullback (1974). For each item, a contingency table is developed which includes the item distractors and the culture groups as is done by Veale and Foreman. However, Maw includes additional variables which are known correlates of educational achievement; e.g., home background and attitudes, instructional processes, and socioeconomic status. These known correlates are expected to account for most of the item variance. Various loglinear models are fitted to the data until the data is

adequately represented. The parameters of the model are then investigated for information about the distractor patterns. Biased items are identified by a significant distractor-by-culture group marginal effect. The individual parameters of the marginal are then analyzed to determine the contributions of the various item response choices.

Item Characteristic Curve Theory Approaches

Recently, latent trait theory has been used to identify biased items (Green and Draper, 1972; Lord, 1977; Rudner, 1977a; Pine, 1976; Scheuneman, 1975, 1976; Durovic, 1975; Wright, Mead and Draba, 1976). In an early study, Green and Draper had used observed total scores as estimates of examinees' abilities (θ_i's) and the proportions of examinees responding correctly at each total score level as estimates of the probability of a correct response given θ_i $[P(u_g = 1 \mid \theta_i)]$. Their procedure called for plotting item character estimated iche curves (icc's) for each item, separately for each culture group, and comparing the plots.

By this and other latent trait theory approaches, an item is unbiased if examinees of the same ability level, but of different cultural affiliations, have equal probabilities of responding correctly. That is, an item is unbiased if the estimated icc's obtained from the various culture groups are identical. As an example of a biased item, consider the two hypothetical curves shown in Figure 1. These curves are based on responses by two different culture groups to the same item. Total observed scores are used as estimates at θ_i and proportions

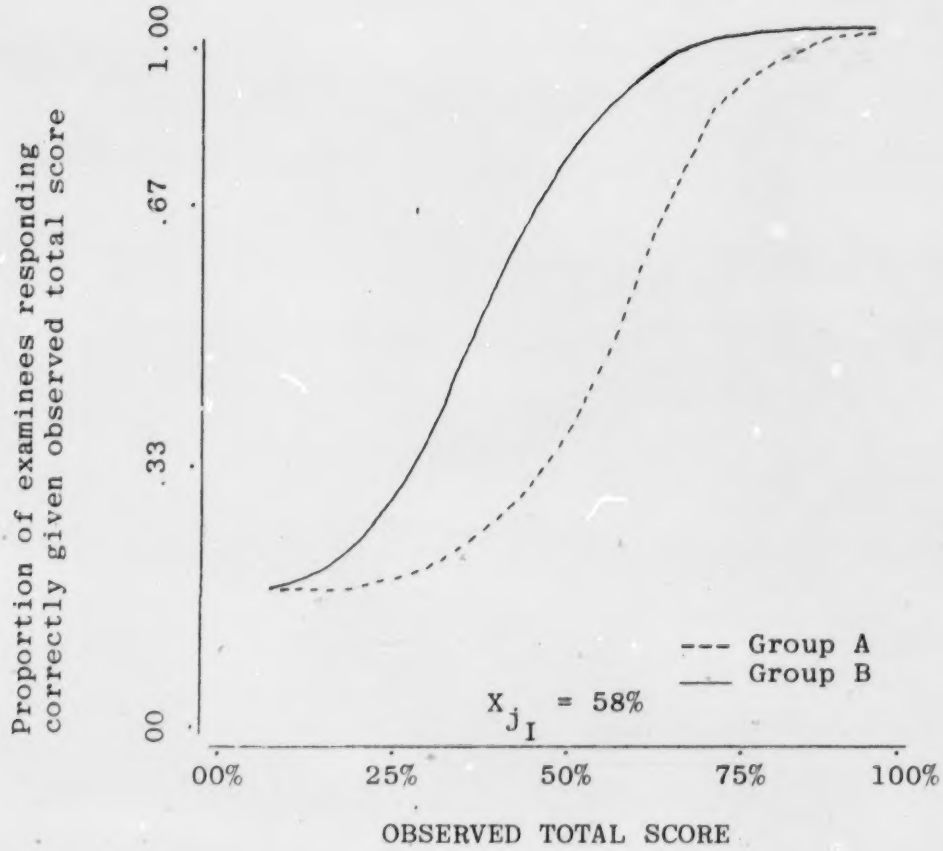

Figure 1: Two hypothetical response distributions

of examinees responding correctly are used as estimates of $P(u_g = 1 \mid \theta_i)$. The curves are not identical, since the location parameters for the two curves are not equal. Such an item can be considered biased in that often examinees of the same ability level; e.g., $X_j = 59\%$, but from different culture groups, do not have similar proportions of correct responses.

While this approach is appealing, total observed scores are directly incorporated and quantification of the degree of item bias is difficult (an eyeballing procedure is used to identify a "very biased item").

Rather than using total observed scores as estimates of θ_i and proportions as estimates for $P(u_g = 1 \mid \theta_i)$, more accurate values can be obtained using one of the recent methods of parameterization (Urry, 1975; Wingersky and Lord, 1973). During parameterization, the metric used for the θ scale is defined by the ability variance in the examined sample. In order to compare parameters obtained from two different examinee groups, the obtained values must be equated. For the three parameter model, Lord and Novick (1974, Chapter 16.11) and Rudner (1977b) have shown that this can be accomplished by computing the regressions of the parameter values based on one group of examinees on the parameter values based on the other group of examinees.

Rudner (1977a), Lord (1977), and Pine (1976) have refined the procedure used by Green and Draper to identify biased items by incorporating equated icc parameter values for the 3 parameter Birnbaum (1968 model). Rudner used the area between pairs

of equated icc's to indicate the relative amount of aberrance of each item and eyeballing of the equated icc's to provide additional information as to the nature of the abberance. Lord has employed an asymptotoic significance test based on the summed variance-covariance matrices of the equated parameter estimates to test for significant differences between pairs of equated icc's. Pine uses the residuals from equating the difficulty and distrimination parameters as an index of aberrance.

Using the one parameter Rasch model, Durovic (1975) and Wright, Mead and Draba (1976) focus their attention on differences in the relative easiness of the items. The differences between observed item responses and the predicted probabilities of a correct response are computed. Goodness-of-fit residual is then analyzed for between-group differences.

Scheuneman (1975, 1976) has developed a technique which is similar to the multi-parameter item characteristic curve theory approach used by Green and Draper. Coined the Chi-Square approach, this approach seeks to determine whether examinees of the same ability level have the same probability of a correct response regardless of cultural affiliation. This is accomplished by blocking each tryout sample into 3 to 5 groups based on the observed scores and comparing the proportions of students within each level responding correctly. An item is considered unbiased if, for all individuals in the same total score interval, the proportion of correct responses is the same for both groups under consideration.

A modified chi-square is used to estimate the probability

that the item is unbiased by the above definition. The expected values for each cell (E_{ij}) are obtained by multiplying (1) the proportion of all examinees with total scores within interval j responding correctly to the item by (2) the number of examinees within the cell. That is,

$$E_{ij} = \frac{O_{.j}}{N_{.j}} N_{ij}$$

where $O_{.j}$ is the number of examinees in total score interval j responding correctly

$N_{.j}$ is the total number of examinees in Group i and score interval j

As with a conventional chi-square, observed cell values are simply the number of examinees within the cell responding correctly to the item.

Summary

Selected methods for examining the test performance of members of identifiable groups for fairness were presented. Two sets of methodologies were identified: one in which an intact test is administered to members of different groups to provide data for selection; the other, in which items from a pool are examined for systematic differentiation among groups. The purpose of this paper was simply to describe the methods. No attempt to evaluate them was made.

References

Anastasi, A. Psychological Testing (4th ed.). New York: Mac-Millan, 1976.

Angoff, W. H. A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, May 1972.

Angoff, W. H., & Ford, S. F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-105.

Angoff, W. H., & Modu, C. C. Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test. New York: College Entrance Examination Board, 1973.

Cardall, C., & Coffman, W. R. A method for comparing performance of different groups on the items in a test. (RM 64-61) Princeton: Educational Testing Service, 1964.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novik, Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley, 1968, Chapters 17-20.

Cleary, T. A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 115-124.

Cleary, T. A., & Hilton, T. L. An investigation into item bias. Educational and Psychological Measurement, 1968, 8, 61-75.

Cole, N. S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Darlington, R. B. Another look at "culture fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Durovic, J. Definitions of test bias: A taxonomy and an illustration of an alternative model. Unpublished doctoral dissertation, State University of New York at Albany, 1975.

Eagle, N., & Harris, A. S. Interaction of race and test on reading performance scores. Journal of Educational Measurement, 1969, 6, 131-135.

Echternacht, G. A quick method for determining test bias. Educational and Psychological Measurement, 1974, 34, 271-280.

Einhorn, H. J., & Bass, A. R. Methodological considerations relevant to discrimination in employment testing. Psychological Bulletin, 1971, 75, 261-269.

Green, D. R. Reducing bias in achievement tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Green, D. R., & Draper, J. F. Exploratory Studies of Bias in Achievement Tests. Monterey: CTB/McGraw-Hill, 1972.

Gross, A. L., & Su, W. Defining a "fair" or "unbiased" selection model: A question of utilities. Journal of Applied Psychology, 1975, 60, 345-351.

Guilford, J. P. Psychometric Methods. New York: McGraw Hill, 1954.

Hicks, M. M., Donlon, T. F., & Wallmark, M. M. Sex differences in item responses on the Graduate Record Examination. Paper presented at the annual meeting of the National Council of

Measurement in Education, San Francisco, April 1976.

Hoepfner, R., & Strickland, G. P. Investigating Test Bias. Los Angeles: Center for the Study of Evaluation, University of California, 1972.

Jensen, A. P. An examination of culture bias in the Wonderlic Personnel Test. Arlington, VA: ERIC Clearinghouse, 1973 (ERIC Document Reproduction Service ED 086 726).

Ku, H. H., & Kullbach, S. Loglinear models in contingency table analysis. The American Statistician, 1974, 28, 115-122.

Linn, R. L. Fair test use in selection. Review of Educational Research, 1973, 43, 139-161.

Lord, F. M. A study of item bias using item characteristic curve theory. Proceedings of the Third Congress of Cross Cultured Psychology, Tilburg, Holland, 1977.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores (2nd ed.). Reading, MA: Addison-Wesley, 1974.

Maw, C. Item bias and information in item responses. Paper presented at the Psychometric Society Meeting, June 1977.

Merz, W. R. Estimating bias in test items utilizing principal components analysis and the general linear solution. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Merz, W. R. Test fairness and test bias: A review of procedures. Paper presented at the USOE Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Program Evaluation, Reston, VA, May 1976.

Ozenne, D. G., Van Gelder, N. C., & Cohen, A. J. Energizing School Aid Act (ESAT) National Evaluation, Achievement Test Restandardization. Santa Monica, CA: Systems Development Corporation, 1974.

Petersen, N. S., & Novick, M. R. An evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, 3-29.

Pine, S. M. Applications of item characteristic curve theory to the problems of test bias. In D. J. Weiss (Ed.) Applications of Computerized Adaptive Testing (RR 77-1). Minneapolis: University of Minnesota Psychometric Methods Program, March 1977.

Plake, B. S., & Hoover, H. D. An analytical method of identifying biased test items. Paper presented at the annual meeting of the American Educational Research Association. New York. April, 1977.

Rudner, L. M. An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977a.

Rudner, L. M. A closer look at latent trait parameters invariance. Paper presented at the annual meeting of the New England Educational Research Organization, Manchester, NH, May 1977b.

Rudner, L. M. An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation, The Catholic University of America, 1977c.

Rudner, L. M. Using standard tests with the hearing impaired: The problem of item bias. Volta Review, 1978, 80, 31-40.

Scheuneman, J. A new method of assessing bias in test items. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April 1975.

Scheuneman, J. A procedure for evaluating item bias in the absence of an outside criterion. Paper presented at the annual meeting of the American Educational Research Association, April 1976.

Strassberg-Rosenberg, B., & Donlon, T. F. Context influences on sex differences in performance and aptitude tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, D.C., 1975.

Thorndike, R. L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, 63-70.

Thurstone, L. L. A method of scaling psychological and educational tests. Journal of Education and Psychology, 1925, 16, 433-451.

Tucker, L. R. An interbattery method of factor analysis. Psychometrika, 1958, 23, 111-136.

Urry, V. W. Ancillary estimators for the parameters of mental test models. Paper presented at the American Psychological Association Convention, Chicago, August 1975.

Veale, J. R., & Foreman, D. I. Cultural validity of items and tests: A new approach. Score Technical Report, Iowa City, Iowa: Westinghouse Learning Corporation/Measurement Research Center, 1975.

CALIFORNIA STATE UNIVERSITY, SACRAMENTO
PROGRAM DOCUMENTATION

USER MESSAGES

Program No.: RGS121

The user may be Job Control or any campus agency. Explain the necessary actions associated with all printed messages.

MESSAGE	ACTION
XXXX MEETS AFTER 10 PM	WARNING TO ACADEMIC ADMIN.
EXAMPLE: XXXX ANTH 010 XXXX ANTH110 SAME CLASS	SHOULD BE CHECK OUT BY ACAD. ADMIN. CORRECT IT IF NOT REALLY TRUE.
THERE IS A TIME CONFLICT BETWEEN XXXX AND XXXX.	THE SAME INSTRUCTOR IS LISTED AS TEACHING 2 CLASSES AT ONCE. CORRECT RHOUGH RGS127.
'X' IS AN INVALID DAY OF THE WEEK COURSE XXXX.	CORRECT THROUGH RGS127

Procedure RGS121 - Input	
Source Document ROOM RESER	
Document Source	

EYPUNCH INSTRUCTIONS	Operation No.
	Procedure No.
ATIONS	
	CARD FORM AND NUMBER

	Card Field	Alpha/ Numeric	No. of Cols.	Verify	Description
1	1- 8	--	8	v	Blank
2	9-13	A	5	v	Days of the week (Required)
3	14-17	N	4	v	Starting time (required)
4	18-21	N	4	v	Ending time (required)
5	22-23	N	2	v	Building code (required)
6	24-28	A/N	5	v	Room number (required)
7	29-38	A	10	v	Instructor's name
8	39	N	1	v	School code (required)
9	40-44	A	5	v	Department name (required)
10	45-49	A/N	5	v	Course Number
11	50-80	--	31	v	Blank
12					
13					
14					
15					
16					
17					

Documents to Next Operation ☐ Other

Cards to Next Operation ☐ Other

Prepared by Clyde M. King Date 01/2/74

30

Data Pr IX-28
ocessing EAM Operating Procedures